Philipps Universität Marburg

Chaos Game Representation für mikrobielle Ökologie von europäischen Seen

Theodor Sperlea (AG Bioinformatik, Prof. Heider)

theodor.sperlea@uni-marburg.de

Bachelor- oder Masterarbeit in Bioinformatik (Machine Learning, Feature Engineering, Ökologie)

Eine der wichtigsten Schritte in *machine learning*-Workflows ist die Wahl einer Darstellung der Daten (*feature encoding*). Dies ist besonders wichtig, wenn es sich bei den Daten um DNA-Sequenzen handelt, da diese üblicherweise als Zeichenketten gespeicherten Informationen zunächst in Zahlen umgewandelt werden müssen. Eine der meistversprechenden *feature encodings* ist die Chaos Game Representation (CGR) [1, 2]. Diese stellt einzelne DNA-Sequenzen als fraktale, zweidimensionale Bilder dar. Allerdings ist es auch möglich, durch Überlagerungen mehrere Sequenzen in einer Abbildung darzustellen.

In einem Projekt in unserer Arbeitsgruppe beschäftigen wir uns mit der Frage, ob sich an Hand der Mikroorganismen, die in einem See leben, Aussagen über dessen Wasserqualität getroffen werden können. Die Anzahlen der Mikroorganismen werden dabei über DNA-Sequenzierung von Wasserproben gewonnen. Bisher arbeiten wir hierbei nur mit den aus diesen DNA-Sequenzen gewonnen Organismenzahlen, wollen nun aber auch andere *feature encoding*-Methoden anwenden.

In dieser Abschlussarbeit soll bestimmt werden, ob sich CGR als feature encoding für ökologische Fragestellungen eignet. Dafür soll die im R Paket kaos implementierte Methode zur Erzeugung von CGRs auf größere DNA-Datensätze angewendet werden und die dadurch entstehenden Encodings im Hinblick auf die Vorhersagekraft einiger ökologischen Parameter durch verschiedene *machine learning*-Ansätzen bewertet werden. Außerdem soll eine Methode entwickelt werden, mit der von vorhersagerelevanten Pixeln der CGRs auf die darunterliegenden DNA-Sequenzen rückgeschlossen werden können.

Literatur

[1] H. J. Jeffrey: Chaos game representation of gene structure. Nucleic Acids Res. 1990, 18(8):213322100.

[2] H. F. Löchel, D. Eger, T. Sperlea, D. Heider: Deep Learning on Chaos Game Representation for Proteins. Bioinformatics 2019, btz493.

Bitte melden Sie sich bei Interesse per Email bei: **30.08.19**

Theodor Sperlea - theodor.sperlea@uni-marburg.de